APPLICATION UNDER UNITED STATES PATENT LAWS

Atty. Dkt. No. PW 283207
(M#)

Invention: System and Method for Managing and Presenting Information Derived from Gene Expression Profiling

Inventor (s): ROGERS, John C.

Pillsbury Winthrop LLP
Intellectual Property Group
1600 Tysons Boulevard
McLean, VA 22102
Attorneys
Telephone: (703) 905-2146

This is a:

- ☐ Provisional Application
- ☐ Regular Utility Application
- ☐ Continuing Application
  - ☒ The contents of the parent are incorporated by reference
- ☒ PCT National Phase Application
- ☐ Design Application
- ☐ Reissue Application
- ☐ Plant Application
- ☐ Substitute Specification
  Sub. Spec Filed _____
  in App. No. ___/_____
- ☐ Marked up Specification re
  Sub. Spec. filed _____
  In App. No ___/_____

SPECIFICATION

SYSTEM AND METHOD FOR MANAGING AND PRESENTING INFORMATION DERIVED FROM GENE EXPRESSION PROFILING

BACKGROUND OF THE INVENTION

1. Copyright Notice.

2. Field of the Invention.

The present invention is directed to certain systems and methods for managing and presenting information derived from techniques for monitoring differential expression of nucleic acid sequences, e.g., gene expression profiling.

3. Description of Background Information.

Gene expression profiling processes are commonly used to represent a cell's physiological response to a particular compound, treatment, or disease. For example, a January 1, 1999 article by Iyer et al., Volume 283, Science, at pages 83-87 (www.sciencemag.org), discloses the use of a temporal program of gene expression to represent a physiological response of human cells to a treatment -- particularly, the response of fibroblasts to serum. A cDNA microarray was used, representing over 8,600 distinct human genes. Fibroblasts, cultured from human neonatal foreskin, were placed in a quiescent state by depriving the cells of serum for 48 hours. The fibroblasts were then stimulated by adding a medium containing 10% FBS, and the microarray was then used to measure the levels of 8,613 different mRNA sequences at 12 distinct times. The microarray was used to identify those genes (including expressed sequence tags -- ESTs) which were substantially repressed or induced and the extent of repression or induction (i.e., fold change). Five hundred seventeen genes whose mRNA levels changed in response to the treatment were selected, and graphically depicted in accordance with a hierarchy.

From this information, various proteins could be identified, which were categorized according to their biological functions. Those biological function categories identified were

signal transduction, intermediate-early transcription factors, other transcription factors, cell cycle and proliferation, coagulation and hemostasis, inflammation, angiogenesis, tissue remodeling, cytoskeletal reorganization, re-epithelialization, cholesterol biosynthesis, and an unidentified role in wound healing.

Various technologies are available for expressing large numbers of genes. A small sample of the available implementations incorporating those technologies include SAGE (serial analysis of gene expression), oligo arrays, and cDNA arrays.

Those technologies produce data identifying large numbers of expressed genes, and the extent of their repression or induction. To aid in the analysis of these large sets of data, biological computational analysis systems are being developed. An approach typically used to create control and treatment probes comprising respective arrays is one used by Iyer et al., in which the data from such arrays is presented in the form of a two-dimensional cluster image showing the dispersion of gene clusters that are either up or down regulated (induced or repressed).

Databases and wall charts have been provided which facilitate the study of treatment data. For example, the Boehringer Mannheim biochemical pathways wall chart and the Cell Signaling Pathways Chart, distributed by Zymed Laboratories, graphically illustrate select metabolic pathways existing in nature, the interrelationships between various of the illustrated metabolic pathways (such as connections between the metabolic pathways, and branching points of substrate metabolism), and factors controlling the direction and the speed of turnover from one point to another within a given metabolic pathway.

There is a need for a system which will better facilitate the analysis of data obtained from expression profiling techniques, to more readily identify key metabolic pathway information, mechanisms of action, mechanisms of drug inactivation and clearance, and potential side effects. Such a system will preferably also provide meaningful information that assists with the identification of the physiological affects of certain treatments and the biological function associated with the affected metabolic activity.

4. Definitions

For purposes of clarification, and to assist readers in an understanding of the present invention, and the embodiments disclosed herein, a number of terms used herein are defined as follows:

Biological function:

an inferred functional classification of a given gene, protein, nucleic acid sequence, or pathway. Some examples of biological functions are metabolism, angiogenesis, signal transduction, transcription factors, cell cycle control, regulation of proliferation, coagulation and hemostasis, inflammation, and apoptosis.

Enzyme:

Protein that catalyzes biochemical reactions.

Protein molecule:

One or several polypetide chains of amino acids.

Expression profiling:

A process by which gene expression techniques are used to measure and compare levels of certain nucleic acid sequences (e.g., mRNAs, proteins, genes, ESTs) in a cell-derived sample in relation to the levels of the same nucleic acid sequences from a different sample or from the same sample at a different time.

Gene:

A sequence of nucleotides specifying a particular polypeptide chain.

Metabolic pathway:

Any individual biological reaction involving a substrate and a product caused by a reaction, as well as the catalyst of such reaction. Catalysts of reactions in metabolic pathways are typically enzymatic. A metabolic pathway also includes any related series of such individual reactions.

Mechanism of action:

A causal link between a variant and a response to the variant, for example, identifying which specific, or where within an individual, metabolic pathway or biological function does a compound or treatment act to produce a given physiological effect. For example, if blood pressure is reduced, the mechanisms of action comprise the specific metabolic pathways and biological functions are being acted upon or involved with the reduction of blood pressure.

mRNA (messenger RNA):

An RNA molecule synthesized from a DNA template -- by the enzyme RNA polymerase. An mRNA functions as a template for the assembly of a polypeptide chain, a process known as translation.

Physiological affect:

Some physiological change or response. A physiological affect could be a state of a given biological system (activation or deactivation), for example a change in high blood pressure.

RNA:

Ribonucleic acid.

RNA Polymerase:

An enzyme that synthesizes RNA by using DNA as a template.

Transcription:

A process by which an RNA molecule is synthesized by the enzyme RNA polymerase using DNA as a template.

SUMMARY OF THE INVENTION

In view of the above, the present invention, through one or more of its various aspects and/or embodiments, is thus presented to accomplish one or more objects and advantages such as those noted below.

An object of the present invention is to provide an improved mechanism for facilitating the display of meaningful information based upon expression profiling, such information facilitating the determination of biological functions involved with treatments, compounds, or diseases, the identification of metabolic pathways, and the identification of mechanisms of action. A further object of the present invention is to provide a structure for organizing and displaying information to enable data mining, whereby expression profile data is grouped in accordance with certain metabolic pathway characteristics in a displayed map.

The present invention, therefore, is directed to a system or method, or one or more components thereof, for managing and presenting information derived from differential expression of genetic information which can be to model a physiological response of biological cells. The system comprises an expression profiling subsystem. The expression profiling subsystem generates, from control and treatment sets of cell-derived samples, respective sets of sequence data representing a direction and a magnitude of regulation of

each one of a high number of different nucleic acid sequences. Sets of nucleic acid sequences are associated with particular regions on a map of metabolic pathways of the biological organism being studied. An overview of the map coordinates may be provided, and those areas or regions of the map comprising high concentrations of affected nucleic acid sequences may be differentiated from other regions of the map, for example, by having a different color. Regions of the map with high concentrations of the affected nucleic acid sequences may be viewed in further detail, to view the specific metabolic pathways involved, and the role the affected nucleic acid sequences play within such metabolic pathways.

Alternatively or in addition, an overview may be provided of the map which identifies specific affected nucleic acid sequences within a given set of metabolic pathways, such indications include a first symbol representing a point of inhibition within the set of pathways, second symbols representing biological catalyst locations within the set of pathways, and third symbols representing locations of end products of the illustrated set of metabolic pathways.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is further described in the detailed description which follows, by reference to a noted plurality of drawings, by way of non-limiting exemplary embodiments of the present invention, in which like reference numerals represent similar parts throughout the several views of the drawings and wherein:

Fig. 1 is a block diagram of a gene expression profiling data analysis system;

Fig. 2 is a flow diagram of a gene expression profiling process;

Fig. 3 is a flow diagram of a process for managing information derived from gene expression profiling;

Fig. 4 is an overview representation of a biochemical pathway map, indicating the concentrations of affected nucleic acid sequences at certain coordinates within the map;

Fig. 5 is a more detailed blown-up view of certain cells within a given area of the biochemical pathway map;

Fig. 6 shows a given set of biosynthetic pathways affected by feedback inhibition;

Fig. 7 is a diagram of a database structure in accordance with the illustrated embodiment;

Fig. 8 is a flow chart representing a process performed by the client computer to match expression profiling data up with mapped metabolic pathways;

Fig. 9 is an example of an overview display of the metabolic pathway map in which related repressed and induced biological catalysts, a point of inhibition and end products are represented by symbols; and

Fig. 10 is a flow diagram of a process of identifying BCIs within affected pathways with a simplified set of symbols on the overview map display.

DETAILED DESCRIPTION OF THE EXEMPLARY EMBODIMENTS

Referring now to the drawings in greater detail, Fig. 1 shows an analysis system 10 according to the illustrated embodiment of the present invention. An expression profiling subsystem 12 is provided, which is coupled to a client computer 14. Client computer 14 comprises, among other elements, a browser application 16, a human interface 18, and a display 20. Human interface 18 may comprise any standard or other interface for facilitating human interaction with and control of client computer 14, including, for example, a keyboard and a mouse. Client computer 14 is coupled to a host computer 24 via a network connection illustrated in Fig. 1 as an intranet. Host computer 24 is connected to a database 26.

Expression profiling system 12 may comprise, for example, an Affymetrix cDNA array. It generates, from control and treatment sets of cell-derived samples, respective sets of sequence data representing a direction and a magnitude of regulation of each one of a high number of different nucleic acid sequences.

Client computer 14, together with human interface 18, display 20, and browser application 16, allows a user to operate analysis system 10. Client computer 14 communicates with database 26 through intranet 22 and host computer 24. Expression profiling subsystem 12 obtains the expression profiling data and stores that data in an organized fashion on database 26.

Host computer 24 is provided with, among other elements, an analysis application 27 for carrying out certain analysis process steps associated with expression profiling and managing the data acquired from the expression profiling. A database server software component 28 is provided for handling and acting on database queries and responses.

Fig. 2 generally shows an expression profiling process in accordance with the illustrated embodiment. In an initial step S2, sequences are generated based upon a baseline

sample (otherwise referred to as a control sample) of cells. One or more differentiated sequences may be generated based upon treated samples, i.e., samples of cells based upon those cells entering into a diseased state or being treated with a particular compound. After performing each of respective steps S2 and S6, a cluster algorithm S4 and S8 is performed, in which similar sequences, including expressed sequence tags (ESTs) are grouped together. Clustering of gene sequence pieces allows redundancies to be eliminated, as a gene expression array will typically identify not only full gene sequences or full mRNA, but will also identify ESTs, which comprise shorter pieces of the full sequence. The total number of sequence pieces within a given cluster may be considered to represent the total number of genes repressed or induced having a particular sequence.

An alternative method of clustering is to use the expression data to cluster by expression patterns, i.e., similar profiles over a course of time. This approach would allow comparison between genes having known functions with genes having unknown functions to assist in identifying the unknown functions, such as is done by Iyer et al. in the above-identified article.

In order to determine whether the gene clusters have been substantially affected (i.e., either repressed or induced), the number of genes generated in the baseline sample is compared with the number of genes generated and clustered in each cluster in the treated sample or samples, to produce, for each treated sample, an indication of whether the gene cluster was regulated and the extent and direction of that regulation.

More specifically, by way of example, a sample of cells may be sequenced using an expression profiling array, such as an Affymetrix GeneChip™ probe array for, for example, the human genome, which is capable of detecting over 6,000 sequences for that genome. Affymetrix provides a GeneChip™ fluidics station which automates the hybridization of nucleic acid targets to a probe array cartridge, and thus controls the delivery of reagents and the timing and temperature for hybridization. Each fluidics station can independently process four probe arrays at a given time.

Accordingly, each target may be prepared from a set of cell dishes by isolation of RNA over a course of time. The treatment of those cells may be emulated by adding, for example, serum thereto. At predetermined intervals, a small amount of the fluid is removed, and the cells are put in a quiescent state to stop the reaction time. Accordingly, a large set of targets, having a predetermined amount of liquid (e.g., .5 ml each) is produced. The

GeneChip™ fluidics station will then automatically hybridize each target, i.e., it will extract all the RNA and label the RNA by adding a chemical tag to each molecule, and control the delivery of the resulting liquid to the probe arrays to facilitate the obtaining of sequencing information regarding the mRNAs. This is done by the probe arrays exposing the target to light at a predetermined location and measuring the photons collected at various locations within the arrays. The amount of mRNA (or an EST) is then ascertained based upon the signal strength of the reading given by the probe at the appropriate location corresponding to that sequence or sequence segment.

Fig. 3 is a flowchart of an analysis process performed by the illustrated embodiment. In a first step S20, gene expression profiling is performed, at which time respective sets of sequence data are generated from control and treatment sets of cell-derived samples/targets, and the obtained data includes information regarding the direction and magnitude of regulation of each one of a high number of different nucleic acid sequence clusters. Once gene expression profiling is performed at step S20, a set of data D2 is produced which comprises the identified sequences and associated regulation information. Then, at step S22, each sequence cluster is matched to a biological catalyst identifier (BCI). In the illustrated embodiment, the BCI may comprise, for example, an EC number. EC numbers are part of a known system for enzyme classification. Each EC number comprises a first number which refers to one of six main subdivisions, a second number which indicates a subclass, a third number indicating a sub-subclass, and a fourth number which represents a serial number. The major EC classes include (1) oxodoreductases-redox reactions (2) transferases -- transfer a group (CH_3), (3) hydrolases -- cleavage, H_2O, (4) lyases -- cleavage by elimination, (5) isomerases -- geometric changes, and (6) ligases -- coupled to ATP hydrolysis. As an example of some subclasses, the oxoreductases are as follows:

1. Oxoreductases

1.1. CHOH doners

1.1.1 NAD^+ or $NADP^+$ acceptor

1.1.2 Cytochrome acceptor

1.1.3. Oxygen acceptor

1.1.5 Quinone acceptor

1.1.99 Other acceptor

At step S24, each cluster of affected sequences, i.e., sequences that have been significantly regulated (by at least twofold) is categorized in accordance with its cluster, whether it was up or down regulated (i.e., induced or repressed, respectively) and the extent of regulation, and further, the number of regulated sequences or sequence segments (ESTs) falling within a given cell of contour plot 30 is summed and binned in association with that cell. This is performed at step S26.

At step S28, a summed section of a detailed map view is displayed, which includes metabolic pathways corresponding to substantially affected sequences. Fig. 4 illustrates a contour plot view 30 of a biochemical pathway map, which illustrates and helps clarify the acts performed in steps S24 and S26 of the process illustrated in Fig. 3.

More specifically, Fig. 4 is a representation of a contour plot view of a biochemical pathway map. In the illustrated embodiment, the map corresponds to the biochemical pathways wall chart by Boehringer Mannheim. The map may comprise graphic representations of biochemical pathways which are identical or comparable to the Boehringer Mannheim wall chart, or any other appropriate set of graphical representations of biochemical pathways, where a given pathway, or point within a pathway, is associated with a particular set of coordinates within the map. In the illustrated embodiment, a matrix of cells is provided, comprising fourteen columns along the X direction (X1-X14) and eight rows along the Y direction (Y1-Y8). The contour plot view shown in Fig. 4 shows whether the number of sequences having an EC number within a given cell is within one of five prescribed ranges. Those ranges are depicted by a different pattern, and include 0,1, 2-3, 4-5, 6-7, and 8-10. By way of example, the cell at coordinates X8, Y7 has one sequence having an EC number falling within that cell. The cell at X7, Y7 has seven sequences with an EC number falling within that cell. Accordingly, the cell X8, Y7 is illustrated as falling within the second range, 1, and the cell at X7, Y7 is shown as having a number of sequences falling within the range 6-7.

While patterns are shown in Fig. 4 in order to differentiate between different ranges of sequences having an EC number falling within a given cell, it is preferred that the ranges be depicted with the use of a coloring scheme. By way of example, the range 1 could be represented by the color purple, while the range 2-3 is represented by the color green, the range 4-5 is represented by the color yellow, the range 6-7 is represented by the color orange, and the range 8-10 is represented by the color red.

The view provided by the contour plot shown in Fig. 4 can thus provide a quick overall view of the activity throughout the various areas of the pathway map, and those areas having yellow, orange and red colors indicate those areas with the most activity. Accordingly, one can select areas in accordance with the amount of activity to view a more detailed view of the map.

Fig. 5 shows a small portion of a biochemical pathway map which illustrates various aspects of certain biochemical pathways at prescribed coordinates x_{n-1}, x_n and x_{n+1} along the x direction, and y_{m-1} , y_m, and y_{m+1} along the y direction. The map comprises graphical representations of metabolic pathways. Those graphical representations comprise individual graphic representations of such items as substrates, products, biological catalysts (BCs), inhibitors, biological functions, and pathway directions (including unique graphical identifiers showing a direction of a pathway in one direction versus the opposite direction, and an amphibolic pathway direction which indicates that the reaction can go in either direction).

More specifically, as shown in Fig. 5, a plurality of pathway direction symbols 40a-40d are provided in the section of the map shown in Fig. 5. The use of an arrow at each end of the illustrated lines 40b and 40c indicates that the pathway direction is amphibolic. A plurality of substrate/product symbols 42a-42c are provided which represent substrate/products$_{1, 2 \text{ and } 3}$. Those symbols may comprise, for example, text identifying a given compound which may serve as either a substrate or a product, depending upon the direction of the chemical reaction. Each biological catalyst or set of biological catalysts associated with the particular pathway, including biological catalyst(s)$_1$ and biological catalyst(s)$_2$ in the illustrated embodiment, is illustrated with a respective biological catalyst symbol 44a,b adjacent to the pathway direction symbol.

A block is provided for indicating a biological catalyst symbol 44a and 44b. These symbols may simply comprise a textual representation of the common nomenclature for the given biological catalyst, which typically will comprise an enzyme in the case of metabolic pathways. BCI (biological catalyst index) symbol 46a, 46b is provided adjacent its respective biological catalyst symbol 44a, 44b, and in the illustrated embodiment simply comprises a numerical representation of the BCI. Any inhibitors will be represented with inhibitor symbols 48a, 48b, which, in the illustrated embodiment, may simply comprise text representing the inhibitor using standard nomenclature.

The biological function with which the metabolic pathways in a certain region of the map are associated may be represented with a biological function symbol 50, which, in the illustrated embodiment, comprises a text representation of the biological function using common nomenclature. Some example biological functions include fatty acid oxidation, carotenoids, and ketone bodies. Other functions include, for example, sulphur metabolism and pterine biosynthesis.

In the illustrated embodiment, one or more of the graphic representations may have a unique color to identify the type of information it is representing. For example, the text serving as BCI symbols 46a, 46b may be in green, the text serving as the biological catalyst symbols 44a, 44b may be magenta or aqua, the text serving as the inhibitor symbols 8a, 48b may be the color brown, and the text serving as the biological function symbol 50 may be the color blue. Additional or alternative coloring schemes may be used. Also, unique graphical patterns may be used in addition or instead of colors to facilitate the viewer's ready identification or classification of a particular symbol as representing one type of information versus another. The enzymes shown in Fig. 5 may have two colors, one if it is induced (up regulated), and another if it is repressed (down regulated). Accordingly, in the illustrated embodiment, biological catalysts 44a and 44b are magenta and aqua, respectively, indicating that biological catalyst $(s)_1$ 44a was induced, while biological catalyst $(s)_2$ 44b was repressed (down regulated).

By mapping sequences obtained from expression profiling techniques to specific symbols within a metabolic pathway map, such as shown in Fig. 5, the information provided by the expression profiling data can be quickly related to meaningful pieces of information relevant to key concerns associated with the treatment, disease, or compound being applied to the tested cells. The visualization of the results of the expression profiling experiment is enabled by identifying such valuable pieces of information as biological function (represented by a biological function symbol 50), metabolic pathway (represented by a set of graphical representations forming a given metabolic pathway at specific coordinates within the map), and a mechanism of action (the identification of which will be more fully described by the use of an example below).

This can have significant benefits in the evaluation of treatments and compounds, for example, allowing the identification of mechanisms of action, mechanisms of drug inactivation and clearance, and potential side effects.

Fig. 6 is an illustration of a select group of related pathways. The related pathways shown in Fig. 6 may correspond, for example, to a number of identified biological catalysts on the map as depicted in the "big picture" view provided in Fig. 9, which will be described further below.

Fig. 6 shows a composite pathway comprising a plurality of pathways ($pathway_1$ - $pathway_9$). Each illustrated pathway ($pathway_1$ - $pathway_9$) may comprise one or more metabolic pathways, as such pathways exist in nature. In this regard, a reference may be made, for example, to the Boehringer Mannheim biochemical pathways wall chart. The specific pathway shown in Fig. 6 can be viewed to identify mechanisms of action, and toxicology and side effects.

Many biochemical pathways involve a long chain of distinct chemical reactions catalyzed by distinct enzymes. The first committed step in a biosynthetic pathway is often regulated by the final product of the pathway through a process called feedback inhibition. Inhibition of a specific enzyme along a metabolic pathway leads to increased levels of intermediate chemicals preceding the point of inhibition, and decreased levels of metabolites following the point of inhibition.

In the composite pathway shown in Fig. 6, a point of inhibition A is shown. Enzymes in the pathway following the point of inhibition A are repressed, while enzymes in another direction following the point of inhibition A are induced. When this occurs, a pathway is inhibited which prohibits the formation of a given final product, and removes any feedback inhibition. Specific enzyme inductions or repressions in response to a disease state, or application of a drug to the system, can be used to identify those pathways which are affected by the disease or drug.

For example, as shown in Fig. 6, a drug may be found to decrease serum cholesterol levels when given to an animal, and that drug may work by an unknown mechanism which is revealed by the graphically-represented pathways. Since cholesterol biosynthesis occurs primarily in the liver, the liver can be removed and mRNA can be isolated therefrom. Using expression profiling techniques, one can determine how this inhibition affects the mRNA level of thousands of enzymes acting in dozens of pathways. The pathways whose enzyme levels are significantly affected by drug treatment indicate the pathway and likely suggest a mechanism of drug action.

This is the case for inhibitors of hydroxy-methyl-glutaryl-CoenzymeA (HMG-CoA) reductase, which is the first step in cholesterol biosynthesis. This step is shown at the top of Fig. 6.

Along pathway$_1$, HMG-CoA is converted to long-chain fatty acids by way of Acetyl-CoA in two reaction steps (not specifically shown in detail in Fig. 6). In another direction, HMG-CoA is converted to a five carbon isoprenoid via a pathway$_4$, and then to a ten carbon geranyl via a pathway$_5$. After another pathway$_6$, a product 15 carbon farnesyl is produced. Another pathway$_7$ produces a 30 carbon squalene, which is then converted to the steroid lanosterol, via pathway$_8$. Then, after pathway$_9$, which comprises a plurality of other reaction steps, cholesterol is produced.

When the drug (HMG-CoA reductase inhibitor) is applied to the liver, and expression profiling is performed on the treated liver, the HMG-CoA reductase and enzymes involved in fatty acid metabolism (which go along the direction of pathway$_1$-pathway$_3$) are induced, and the enzymes involved in the formation of cholesterol are repressed.

The identification of pathways of drug metabolism and elimination is done similarly. Most drugs are metabolized by oxidation to a more reactive species than conjugation to a sugar or other molecule that is recognized in the kidney for elimination. The oxidative step is catalyzed by one or more of over 200 enzymes, including cytochrome P 450 enzymes, followed by conjugation by conjugating enzymes in the liver. These enzymes may be induced directly by the drug, or because the drug competes with a normal substrate, in which case less of the normal product is produced by the enzyme pathway, and feedback by that product is reduced.

Induction of some genes is indicative of toxic effects. A variety of enzymes involved in drug metabolism are induced in tumor cells (P450 4 F1) and the induction by a drug can indicate that a drug is potentially tumorigenic. In addition, metabolism of a drug may create toxic metabolites, and may induce peroxidation and proteolytic cascades, which can indicate that a drug or drug metabolite is causing cell death or damage.

Fig. 7 generally shows, in a block diagram, the structure of the database 26 illustrated in Fig. 1. Database 26 comprises, among other elements, seven tables as illustrated in Fig. 1, including table1 (an experiment), table2 (data), table3 (sequence), table4 (BCI link), table5 (BCI number), table6 (map link), and table7 (coordinate).

The experiment, table1, is populated by expression profiling subsystem 12 at some point in time. It includes experiment identifiers (ExpID) and associated experiment names and experiment conditions. Table2 includes the data obtained from the experiment, including the experiment identification (ExpID) the sequence identification, sequence ID, and the fold-change of each sequence that has been identified as being affected. Table1 is linked with Table2 by means of the variable ExpID. Table2 holds an associated sequence ID and fold-change values in association with each ExpID value. The sequence ID value within table2 is associated with a corresponding indexed sequence ID in table3 which serves as a sequence table. For each sequence ID, additional variables are associated therewith, including an accession variable, and a description of the sequence.

A BCI link table4 is provided which is linked to table2 and table3 in accordance with a sequence ID index thereof. BCI link table4 associates with each sequence ID values including BCI ID, a sequence/link value, and a link score. Each BCI ID has an associated BCI number (BCI) which is listed in table5. Each BCI ID of table4 and of table5 is linked to a BCI ID index provided in a map link table6. Each BCI ID has a coordinate ID associated therewith, which is provided within map link table6. Map link table6 is linked to coordinate table7 by means of a coordinate ID value. Coordinate table7 provides values associated with each coordinate ID value, including an x coordinate of the biochemical pathway map, a y coordinate of the biochemical pathway map, and a biological function associated with the given location on the map per the corresponding x and y coordinates. The database 26 may be implemented, in the illustrated embodiment, in accordance with the third normal form of relational database. It is noted that most of the actual data is stored in table1, table3, table5 and table7, while link tables, table 2, table4 and table6 are provided to primarily minimize redundancy in the database.

Linking tables, table2, table4, and table6, facilitate the many-to-many relationships. Such exist between experiments and genes -- many genes are affected in a given experiment, and many experiments may be done with each gene. There are also many-to-many relationships between genes and BCI numbers (e.g., EC numbers). For example, a multi-functional gene may have many EC numbers, and many similar genes could have the same EC number. Many-to-many relationships also exist between BCI numbers and mapped coordinates. For example, if the BCI number comprises an EC number, and the map comprises or is modeled after the Boehringer Mannheim biochemical pathways wall chart,

one EC number can easily appear more than once within a coordinate or in multiple coordinates, and each coordinate can have many EC numbers.

Fig. 8 is a flowchart illustrating a process of handling data, which is performed by analyzing system 10 in connection with its use of database 26. In a first step S40, experiment data is read and stored in table1. Then, in step S42, the act of storing sequence data in table3 is performed. The experiment data stored in table1 includes, among other data, the experiment ID (ExpID), the experiment name (ExpName), and the conditions of the experiment. The sequence data stored in table3 includes the sequence id, the accession number corresponding to that sequence, and description data concerning the sequence. In step S44, the fold change per sequence (or per sequence cluster) is determined, and that information is stored in table2 and related to other data including ExpId and the sequence ID. In step S46, the BCIs are linked to sequences. Table4 is then used to link the sequences to the BCI data in table5.

In step S48, the BCIs are linked to map coordinates of the map. Link table table6 is used to link the BCIs to the coordinate data in Table7.

Fig. 9 shows another overview display of the map. In this view, a point of inhibition 60 is displayed with a first symbol 60 (which is a square in the illustrated embodiment) at a specific location within a particular cell of the map corresponding to the point in the pathway at which the inhibition occurs. Second symbols 62a - 62l represent enzymes which correspond to sequences affected by the treatment. One color (dark gray in Fig. 9) is used to represent enzymes which are induced, while another color (white in Fig. 9) represents enzymes which were repressed. Third symbols 64a and 64b represent end products of the illustrated pathways. The symbols shown in Fig. 9 are all on a common composite pathway. End product symbol 64a is shown as dark gray because it is the end product of the pathway corresponding to the induced enzymes, while end product symbol 64b is shown as white because it is the end product corresponding to the pathway which is populated by enzymes which were repressed.

The analysis application 27 may be configured so that various display modes are provided, including a first display mode in which the contour map view is provided as shown in Fig. 4, and a second display mode in which respective overview pathways are provided as shown in Fig. 9. When in the second mode, each composite pathway may be separately

illustrated on its own, or one map may be provided on which the unrelated composite pathways are all indicated.

A third display mode may be provided in which a detailed view of the map is provided. This mode may be entered by the user selectively choosing a detailed map at any desired set of coordinates, by simply clicking on the desired coordinates in an overview display in either of the first and second display modes.

Fig. 10 is a flow diagram of those steps performed by analysis application 27 to create the overview display shown in Fig. 9. In a first step S50, the act of determining specific coordinates of BCIs is performed. In a next step S52, the BCIs are determined which are common to the same pathway. If there is more than one separate unrelated composite pathway, a plurality of sets of BCIs are determined and separately categorized. In step S54, the induced BCIs of a given common pathway are displayed, with one color representing induced BCIs and another color representing repressed BCIs.

In step S56, subcoordinates of the point of inhibition are determined -- if there is a point of inhibition, i.e., if one side of the common pathway includes all repressed BCIs, while another side of the common pathway includes all induced BCIs. This point is displayed at the appropriate location within the biochemical pathway map with a second symbol.

At step S58, the subcoordinates of the end products of the common composite pathway are determined, and those points are displayed with a third symbol, with one color representing the end product of a pathway portion corresponding to induced BCIs and another color representing an end product corresponding to the end of a portion of a path corresponding to the repressed BCIs.

The point of inhibition may, for example, be determined by identifying the point along a pathway at which the enzymes switch from one affected state (e.g., induction) to another state (e.g., repression). The end products may, for example, be presumed by determining the point along the pathway at which the enzymes are no longer affected, or with the use of data known about the relevant pathways.

Another, more specific embodiment of the present invention will now be described. This embodiment is merely an illustrative example.

Initially, a database is created which relates EC (enzyme commission) numbers to coordinates on the Boehringer Mannheim biochemical pathways wall chart. This database contains current descriptions for all EC numbers and other information pertaining to the EC

numbers. Descriptions of the EC numbers and other enzyme data are publicly available, and may be obtained from the website http:/www.expasy.ch/txt/enzyme.get. A database may then be created linking the EC numbers with specific map coordinates corresponding to the Boehringer Mannheim biochemical pathways wall chart.

Once expression profiling is performed, and experiment data is obtained, EC numbers are assigned to the sequence clusters obtained in the experiment. This may involve a list of GenBank accession numbers corresponding to those affected genes affected more than two fold in a set of profiling experiments. GenBank records are available at http://www.ncbi.nom.nih.gov/entrez/, and may be parsed for the pattern of numbers in an EC number (#.#.#.#). For every occurrence of an EC number in the GenBank file, a GenBank accession number and corresponding EC number may be written to a text file for loading into a database. The following is a sample GenBank file:

LOCUS 4191746 375 aa 27-JAN-1999
DEFINITION alcohol dehydrogenase; ADH.ACCESSION 4191746PID g4191746
DBSOURCE GENBANK: locus L30113, accession L30113KEYWORDS
SOURCE baboon. ORGANISM Papio hamadryas
Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria;
Primates; Catarrhini; Cercopithecidae; Cercopithecinae; Papio.
REFERENCE 1 (residues 1 to 375)
AUTHORS Cheung,B., Holmes,R.S., Easteal,S. and Beacham,I.R.
TITLE Evolution of Class I Alcohol Dehydrogenase Genes in Catarrhine
Primates: Gene Conversion, Substitution Rates, and Gene Regulation
JOURNAL Mol. Biol. Evol. 16 (1), 23-36 (1999)
FEATURES Location/Qualifiers source 1..375
/organism="Papio hamadryas"
/db_xref="taxon:9557"
/tissue_type="kidney" Protein 1..375
/note="ADH"
/product="alcohol dehydrogenase"
/EC_number="1.1.1.1" CDS 1..375
/note="putative"
/coded_by="L30113:53..1180"ORIGIN

1 mstagkvikc kaavlwevkk pfsieeveva ppkahevrik mvavgicrsd dhvvsgtlvt
61 plpailghea agivegvgeg vttvkpgdkv iplftpqcgk crvcknpesn ycfkndlsnp
121 rgtmqdgtrr ftcggkpihh flgistfsqy tvvdenavak idaasplekv cligcgfstg
181 ygpavkvakv tpgstcavfg lggvglsavm gckaagaari iavdinkdkf akakelgate
241 cinpqdykkp iqevlkemtd ggvdfsfevi grldtimasl lccheacgts vivgvppdsq
301 nlsinpvlll tgrtwkgaif ggfkskesvp klvsdfmakk fsldalitnv lpfekinegf
361 dllrsgksir tilmf//

If no EC number is available in the GenBank file, the nucleotide or amino acid sequence may be obtained from the GenBank file which corresponds to a particular cluster obtained from the expression profiling, and a BLAST sequence alignment may be performed, which may be performed by accessing the publicly available application through http://www.ncbi.nlm.nih.gov/cgi-vin/BLAST/nph-newblast?Jform=0. The GenBank file may then be fetched for each sequence that aligns with an expect value (E value, right-most column in the BLAST results) that is less than 1 e-30, and by looking for EC numbers in these related sequence files. If an EC number is present, the accession number for the gene affected in the expression profiling experiment can be recorded, and the expect value from the sequence alignment may be recorded as well, along with the EC number or numbers found in the related sequence file or files.

At this point, the database can be created, as described previously in this document. In this regard, in accordance with the specific embodiment now being described, database 26 as shown in Fig. 1 may comprise an ORACLE database, and host computer may comprise a Silicon Graphics Origin 2000 computer. These items are merely illustrative, and are not meant to limit the invention in any way. Other computer systems, databases, and database structures may be used.

Analysis application 27 may be implemented with use of a Netscape FastTrack WWW server using standard HTML and Perl. The Perl modules which may be used to implement this application include (1) DBI/DBD -- a database interface for communicating with a remote pathmap database, (2) CGI -- for generating HTML code, (3) PGPLOT -- an interface to compiled PGPLOT Fortran libraries for creating contour plots, (4) GD -- a graphical drawing module for cropping a GIF image produced by PGPLOT and for drawing polygons and rectangles used for background coloring, (5) MLDBM -- a Perl module that

allows creation of a persistent multi-level data structure to implement image map shape data, and (6) ImageMagick -- a module for performing image processing, so that the background created with GD can be used to create masks, overlays and background coloring.

The application may be configured so that a user can connect to a path map web page through the use of browser application 16, select an experiment, and query the database to select the wall chart coordinates of genes affected more than two-fold in the experiment. The number of genes mapped to each map coordinate are binned, and a contour plot of hits per coordinate may be displayed, for example, as shown in Fig. 4. Other displays may be provided, as well, such as those shown in Fig. 9. The user may move the cursor with the use of the mouse to the position on the map image to see the biological function corresponding to that area of the map, and can click on that particular cell of the map to obtain a more detailed view of the pathway information, such as that shown in Fig. 5. In this regard, if the Boehringer Mannheim biochemical pathways wall chart structure is used, it is modified to illustrate the induced and repressed genes, as well as the EC numbers in association with the identified enzymes corresponding to those genes. The enzymes corresponding to affected genes are colored based upon whether the gene was repressed or induced. Specifically, the enzyme may be represented with magenta text if the corresponding gene cluster was induced, cyan if it was repressed, and green if two or more gene clusters with the same EC number were affected in opposite directions. The interface provided to the user through browser application 16 is displayed on display 20, and may provide a mechanism for allowing the user to click on the accession number in order to obtain information on a particular gene and all available experiments pertinent to the gene. A mechanism may also be provided to allow clicking on a particular EC number to obtain all information relating to that EC number. In addition, the analysis system 10 may be provided with a search tool to allow the user to submit queries by any given parameter to obtain information related to that parameter. For example, the user may query by accession number or gene description to find information for a specific gene of interest.